

BayesCAT: Bayesian Co-estimation of Alignment and Tree

Heejung Shim

shim@gmail.com

Department of Human Genetics, University of Chicago

Bret Larget

brlarget@wisc.edu

Departments of Statistics and of Botany, University of Wisconsin, Madison

Abstract

Traditionally, phylogeny and sequence alignment are estimated separately: first estimate a multiple sequence alignment and then infer a phylogeny based on the sequence alignment estimated in the previous step. However, uncertainty in the alignment estimation is ignored, resulting, possibly, in overstated certainty in phylogeny estimates. We develop a joint model for co-estimating phylogeny and sequence alignment which improves estimates from the traditional approach by accounting for uncertainty in the alignment in phylogenetic inferences. Our insertion and deletion (indel) model allows arbitrary-length overlapping indel events and a general distribution for indel fragment size. We employ a Bayesian approach using MCMC to estimate the joint posterior distribution of a phylogenetic tree and a multiple sequence alignment. Our approach has a tree and a complete history of indel events mapped onto the tree as the state space of the Markov Chain while alternative previous approaches have a tree and an alignment. A large state space containing a complete history of indel events makes our MCMC approach more challenging, but it enables us to infer more information about the indel process. The performances of this joint method and traditional sequential methods are compared using simulated data as well as real data. Software named BayesCAT (Bayesian Co-estimation of Alignment and Tree) is available at <https://github.com/heejungshim/BayesCAT>.

1 Introduction

The use of molecular sequences is popular in phylogeny estimation and a large number of statistical methods have been proposed. Phylogenetic inference using molecular sequences traditionally consists of two separate steps: first estimate a multiple sequence alignment

and then infer a phylogeny based on the sequence alignment estimated in the previous step. This sequential approach ignores alignment uncertainty, leading to several problems with phylogenetic inference.

If the alignment contains ambiguous regions, ignoring uncertainty in the alignment can result in exaggerated support for inferred phylogenies (Lutzoni *et al.*, 2000). Moreover, if the sequence alignment is determined by an alignment method that assumes a fixed guide tree, then the estimated phylogeny in the second step may be biased toward this fixed guide tree (Lake, 1991; Thorne and Kishino, 1992; Sinsheimer, 1994; Nelesen *et al.*, 2008). As various alignment methods typically align ambiguous regions differently, phylogenies estimated by the traditional sequential approach can change considerably according to the choice of alignment program. Wong *et al.* (2008) show that different alignment methods can lead to different conclusions in a comparative genomics study. (See Web Appendix A for our investigation of the problems of the traditional sequential approach using a simulated data set). A simple approach to avoid these problems is to exclude ambiguous regions in the following phylogeny estimation procedure. However, the decision of which regions are ambiguous is subjective and ambiguous regions can include a large fraction of potentially informative sites (Lutzoni *et al.*, 2000). Another approach to sidestep the limitations of the traditional sequential approach is to estimate alignment and phylogeny simultaneously. Thus, researchers have developed diverse methods for joint estimation of alignment and phylogeny including statistical approaches (Lunter *et al.*, 2005; Redelings and Suchard, 2005, 2007; Novák *et al.*, 2008) and non-statistical approaches (Varón *et al.*, 2010; Liu *et al.*, 2009, 2012). More comprehensive background on these methods can be found in Redelings and Suchard (2005).

Statistical approaches to joint estimation use mutation rates, insertion and deletion (indel) rates, and divergence time, instead of penalties (e.g., gap and mismatch penalties) employed in most non-statistical methods, so that these approaches take into account multiple occurrences of mutations and indels at each site. In addition, statistical approaches use statistical models for the substitution process and the indel process, allowing for inferences about the nature of the process of evolution. In particular, Bayesian statistical approaches provide a framework to measure uncertainty in the estimated alignment and tree.

Lunter *et al.* (2005) developed a fully Bayesian method which uses the TKF91 model (Thorne *et al.*, 1991) for indel events. The TKF91 model has the restriction of allowing only single-base indels. This restriction tends to overemphasize the information in a single long indel by treating one event as many, which can affect posterior estimates (Redelings and Suchard, 2005). Redelings and Suchard also proposed a Bayesian approach (BALi-Phy). In their first paper (Redelings and Suchard, 2005), they allow indels to contain a geometrically distributed number of bases. Although their model does not allow indels on the same branch to overlap, it improves on the TKF92 model (Thorne *et al.*, 1992) by avoiding a fragment-based indel process. However, this improvement was made possible by assuming that occurrence of indel events on each branch is independent of branch

length. As it is biologically reasonable to expect more indel events on longer branches, this assumption is undesirable. Their second paper (Redelings and Suchard, 2007) removed the assumption, which results in a fragment-based indel model like the TKF92 model on individual branches. Novák *et al.* (2008) also developed a software package where they used the long indel model introduced in Miklós *et al.* (2004). The long indel model improves on the TKF91 and TKF92 models by allowing indels to have multiple bases and overlap. Miklós *et al.* (2004) introduce an algorithm for calculation of alignment likelihoods under the long indel model, but it is based on approximation by bounding the number of indel events and the indel fragment size per event.

The statistical joint estimation methods above sum over all possible indel histories under their models, which yields a restricted inference on the indel process itself. Estimated multiple alignments show inferred homologies, but are not easy to interpret with regard to specific indel event histories. In addition, to achieve this summability, the models disallow many biologically plausible indel histories.

In this paper, we develop a model for joint estimation of alignment and phylogeny and design MCMC methods to carry out Bayesian inference. We propose an indel model which allows arbitrary-length overlapping indel events and a general distribution for indel fragment size. We use the exact likelihood of the indel history under our indel model instead of an approximated likelihood of alignment. The major difference between our approach and the previous approaches to the joint estimation of tree and alignment is the state space of the Markov Chain. Our approach has a tree and a complete history of indel events on the tree as the state space while the previous approaches have a tree and an alignment. A large state space containing a complete history of indel events makes our MCMC approach more challenging, but it enables us to infer more information about the indel process.

2 Model

To model the evolution of molecular sequences, we consider the process of nucleotide substitution in which single sites change bases and the indel process in which DNA fragments are inserted into or deleted from the sequence. In our joint model for co-estimation of phylogeny and alignment, these two processes can be separated, and moreover, the traditional substitution models used with fixed alignments can be adopted. In this paper, we develop an indel model, which allows arbitrary-length overlapping indel events and a general distribution for indel fragment size.

2.1 Joint model

The observed data S consists of n unaligned sequences. The n unaligned sequences are related by a phylogenetic tree T and aligned by a history of indel events H on the tree.

The phylogenetic tree T is composed of an unrooted bifurcating tree topology τ and branch lengths denoted as $V = (v_1, \dots, v_{2n-3})$. The indel history H includes for each edge a sequence of events which consist of the time, type (insertion or deletion), position on the sequence, and inserted or deleted fragment size (see Section 2.3.1 and Section 2.3.5 for details). The unnormalized posterior distribution of the tree T and indel history H given the observed sequences S is:

$$P(H, \tau, V \mid S, \Theta) \propto P(S \mid H, \tau, V, \Theta_{\text{sub}})P(H \mid \tau, V, \Theta_{\text{ID}})P(\tau, V \mid \Theta_{\text{tree}})$$

where Θ consists of three components, Θ_{sub} , Θ_{ID} , and Θ_{tree} , for the nucleotide substitution process, indel process, and the tree, respectively. On the right-hand side of the equation, the first factor is the likelihood of the sequences and is given by a substitution model. The second factor, the probability of the indel history on a given tree is specified by our indel model, which will be described in detail later. For the third factor, we assume a uniform distribution over unrooted tree topologies with n taxa and independent exponential distributions with common mean $1/\gamma$ on the length of each branch, leading to this expression:

$$P(H, \tau, V \mid S, \Theta) \propto P(S \mid H, \tau, V, \Theta_{\text{sub}})P(H \mid \tau, V, \Theta_{\text{ID}})P(\tau)P(V \mid \Theta_{\text{tree}})$$

where $\Theta_{\text{tree}} = \gamma$.

2.2 Substitution model

A history of indel events H on a tree T determines a multiple alignment A (up to minor reordering of some columns) where homologous residues are aligned in columns. Different indel histories might give rise to the same alignment. The alignment A is sufficient for the substitution process, yielding $P(S \mid H, \tau, V, \Theta_{\text{sub}}) = P(S \mid A, \tau, V, \Theta_{\text{sub}})$, where A is a function of H on $T = (\tau, V)$. The traditional substitution models used with fixed alignments can be adopted here since the right-hand side of the equation has the same form. We assume substitutions occur independently across columns of A according to a continuous-time Markov process. At present, we only consider reversible Markov models, which leads to the use of an unrooted tree topology. Any substitution model could be used.

We use the HKY model (Hasegawa *et al.*, 1985) in our analysis here (see Web Appendix B for details of the HKY model), so Θ_{sub} consists of κ , the ratio of the transition to transversion rates among nucleotides, and nucleotide frequencies in the equilibrium distribution of the rate matrix, denoted as $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$.

2.3 Indel model

2.3.1 Description of an indel history on a tree

To help our description of indel models, this section first describes indel events for a given sequence, and then illustrates indel events on a tree using simple examples. A molecular sequence is represented by a sequence of symbols with one symbol for each base in the sequence. We number the bases from left to right in a sequence of length n from one to n . We use the term *position* to refer to the places between or at the ends of bases in a sequence where indel events might act. A sequence of length n has $n + 1$ positions which we number from zero to n from left to right, so base i is between positions $i - 1$ and i . Deletions remove all bases between two positions. We identify a deletion event with the leftmost position and the number of bases deleted. In general, a deletion event of size x at position i removes all bases between positions i and $i + x$. Insertions act at a single position by adding one or more bases to the sequence. See Web Figure 1 for examples of indel events for a given sequence.

We specify a complete history of indel events on a tree relative to rooting the tree at one time point. As depicted in Figure 1, we represent an insertion or deletion event on the rooted tree based on the sequence before the occurrence of that event. Events $E1$ and $E3$ are specified based on the root sequence and event $E2$ is represented relative to the sequence after event $E1$. A history of indel events on a tree determines *homologous* residues, i.e., residues derived from a common ancestor. In Figure 1, residues with the same notation are homologous.

2.3.2 General indel model

We develop a general indel model that allows arbitrary-length overlapping indels and a general distribution for the indel fragment size. We imagine a sequence of interest embedded within a much longer sequence which undergoes a homogeneous process of insertion and deletion, conditional on leaving the endpoints of the sequence of interest intact.

Our indel model rests on the following assumptions: (1) time reversibility; (2) insertions can occur at any positions on a given sequence including the end positions; (3) insertion fragments can be of any size; (4) for a fragment of a given size, the insertion rate is spatially homogeneous on the sequence; (5) deletions can occur at any positions on the sequence except for the end position; (6) for a given position on the sequence, the deletion fragment has maximum size, which is the number of residues to the right of the position; (7) for a fragment of a given size, the deletion rate is spatially homogeneous over possible positions on the sequence; (8) non-zero deletion rate of a single-residue; and (9) the total insertion rate and total deletion rate per site on the sequence are finite.

We specify the constant insertion and deletion rates for a fragment of k bases using $\lambda i(k)$ and $\mu d(k)$, respectively, where $d(j) \geq 0$ and $i(j) \geq 0$ for all $j \in \{1, 2, \dots\}$,

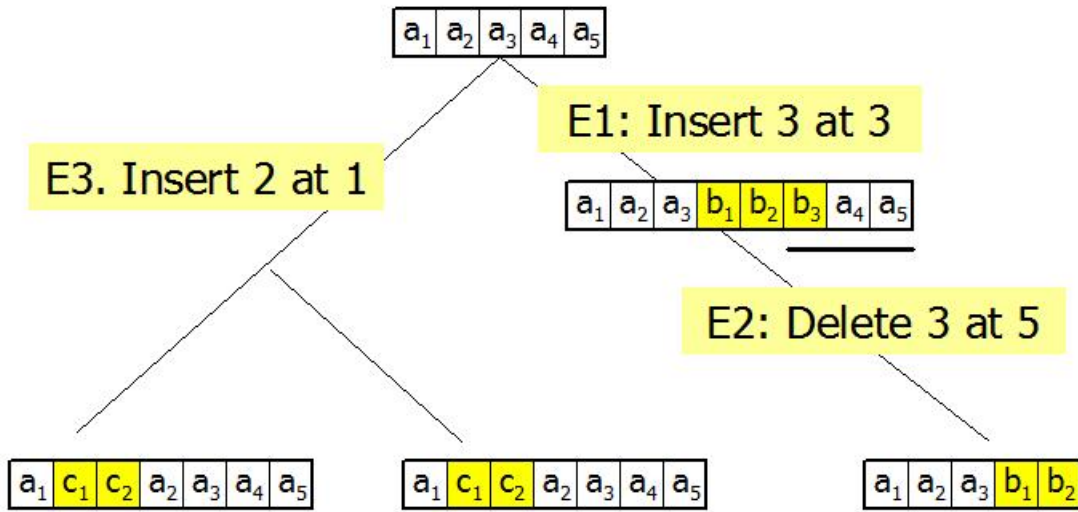


Figure 1: **Example : A complete history of indel events on a tree.** Three indel events occur on a rooted tree, in which a sequence at the root has five residues. Event $E1$ on the right child edge inserts three residues, colored yellow, at position three into the root sequence. Event $E2$, which occurs after $E1$ on the same edge, deletes three residues at position five. The deleted residues are underlined. The rightmost leaf retains five residues. Event $E3$ on the left child edge is independent of $E1$ and $E2$. It inserts two residues at position one in the root sequence. The remaining two leaves have sequences of length seven. Residues with the same notation are homologous.

$\sum_{j=1}^{\infty} i(j) = 1$, and $\sum_{j=1}^{\infty} d(j) = 1$. Then, λ and μ represent the total insertion and deletion rates per site on a sequence of infinite length, respectively. In addition, $i(\cdot)$ and $d(\cdot)$ have the same meaning as the insertion and deletion fragment size distributions on the sequence of infinite length, respectively. In this paper, we call $i(\cdot)$ and $d(\cdot)$ the base insertion and deletion fragment size distributions, respectively.

To clarify these assumptions, we present several comments. First, assumption (6) implies deletions have no possible fragment size at the right end of the sequence, which is the reason we exclude this position from possible positions for deletion in assumption (5). Second, for a given fragment size, deletions have a restriction on possible positions of the sequence due to assumption (6). Third, under the assumptions above, the total insertion rate per site is homogeneous, but the total deletion rate per site depends on the position on the sequence. Due to assumption (6), the total deletion rate per site decreases as the site approaches the right end of the sequence (see the example in Web Figure 2). Fourth, the insertion fragment size distribution is identical to $i(\cdot)$ at all positions. However, due to assumption (6), the deletion fragment size distribution at a given position depends on the position on the sequence and it is truncated distribution of $d(\cdot)$ (see the example in Web Figure 2). Finally, assumption (8) is required in the derivation of our model, but allowing indel events of unit base is also biologically realistic.

The components of a general indel model that follows the previous assumptions include the equilibrium length distribution $q(\cdot)$, base indel fragment size distributions $i(\cdot)$ and $d(\cdot)$, and indel rates λ and μ . The following proposition describes the most general indel model that satisfies the above assumptions.

Proposition 2.1 *Under the previous assumptions, the equilibrium length distribution $q(\cdot)$ is*

$$q(x) = r(1-r)^x \text{ for all } x \in \{0, 1, \dots\},$$

where $1-r = \frac{\lambda i(1)}{\mu d(1)}$ and $0 < r < 1$. The base deletion fragment size distribution $d(\cdot)$ can be any distribution with support on the positive integers and $d(1) > 0$, and the ratio of the insertion rate to the deletion rate is

$$\frac{\lambda}{\mu} = \sum_{k=1}^{\infty} (1-r)^k d(k) < 1.$$

The base insertion fragment size distribution $i(\cdot)$ is determined as

$$i(k) = \frac{\mu}{\lambda} (1-r)^k d(k) \text{ for all } k \in \{1, 2, \dots\}.$$

The proof is given in Web Appendix C. The most general indel model allowed under the assumptions above permits free specification of the parameter r , the distribution $d(\cdot)$, and one of λ or μ , but the remaining components of the model are then determined.

2.3.3 Examples of general indel model

The selection of a particular distribution for the deletion fragment size determines examples of the general indel model. We illustrate a geometric distribution here and negative binomial and power law distributions in Web Appendix D. We consider a geometric distribution with parameter r_d as the deletion fragment size distribution $d(\cdot)$. Then, the insertion fragment size distribution $i(\cdot)$ is also geometric as follows. Selection of $0 < r, r_d < 1$ yields $q(x) = r(1 - r)^x$ for $x = 0, 1, \dots$, $d(k) = r_d(1 - r_d)^{(k-1)}$ and $i(k) = r_i(1 - r_i)^{(k-1)}$ for $k = 1, 2, \dots$, and $\frac{\lambda}{\mu} = \frac{r_d(1-r_i)}{r_i(1-r_d)}$, where $r_i = 1 - (1 - r_d)(1 - r)$, $0 < r, r_d, r_i < 1$, $r_i > r_d$, $r_i > r$, and $\mu > \lambda > 0$. As the choice of $\lambda > 0$ determines μ or vice versa, this model has three free parameters. That is, $\Theta_{\text{ID}} = (r, r_d, \lambda)$. As insertion has one more possible position than deletion on a sequence, the constraint $\mu > \lambda$ is necessary in our model to prevent sequences from growing indefinitely over time. The requirement $r_i > r_d$ is also reasonable because for a given position on the sequence, the deletion fragment size has a maximum possible value while the insertion fragment size is not restricted.

2.3.4 Relationship with the previous indel models

Allowing $r_d = 1$ under the geometric model results in the TKF91 model as a special case of our model.

It turns out that our indel model is very similar to the long indel model (Miklós *et al.*, 2004). Under our indel model, for a fragment of a given size, the insertion rate and the deletion rate per site are spatially homogeneous over possible positions on the sequence. The total insertion rate per site is also homogeneous. However, the total deletion rate per site depends on the position on the sequence because the deletion fragment has maximum size, which is the number of residues on the right hand side of the position. Thus, the total deletion rate per site decreases as the site approaches the right end of the sequence in our indel model. Conversely, the long indel model assumes that the total insertion rate and the total deletion rate per site are spatially homogeneous, which leads to increased insertion and deletion rates of a given fragment size at both ends of the sequence. Miklós *et al.* (2004) introduce an algorithm for calculation of alignment likelihoods under the long indel model, but it is based on approximation by bounding the number of indel events and the indel fragment size per event. In this paper, we use the exact likelihood of the indel history under our indel model instead of an approximated likelihood of alignment.

2.3.5 Specific description of an indel history on a single edge

Although the tree in our model is unrooted, we assume a time direction for convenience when calculating the likelihood or updating an indel history on the tree. Thus, we will describe specific components of an indel history on a single edge and introduce their notation after assuming the single edge has defined parent and child nodes. Let an indel history h on a single edge of length v have K indel events. These events are ordered

by their occurrence time on the edge defined relative to the parent node. In the i th event $e_i = (t_i, id_i, p_i, l_i, n_i)$, t_i indicates the time of this event defined as a distance from the parent node to the event; id_i denotes its type, whether the event is an insertion or deletion; p_i signifies the position on the sequence where a fragment for deletion starts or a new fragment is inserted; l_i is the size of the inserted or deleted fragment; n_i is the total length of the sequence after the i th event. For convenience, let n_0 be the sequence length at the parent node and let $n_{K+1} = n_K$ be the sequence length at the child node. Then, $n_i = n_{i-1} + l_i$ if the i th event is an insertion and $n_i = n_{i-1} - l_i$ if the i th event is a deletion. Let $t_0 = 0$ and $t_{K+1} = v$, which is the length of the single edge.

2.3.6 Indel history probability density calculation

We first derive the likelihood for the history on a single edge and then for the entire tree.

On a single edge Under our indel model, the likelihood for an indel history $h = (e_1, e_2, \dots, e_K)$ on a single edge, conditional on the branch length v and the sequence length of the parent node n_0 , is computed as the product of exponentially-distributed waiting times for each event multiplied by an exponential tail probability for no further events in the remaining interval.

$$P(h \mid v, n_0) = \left[\prod_{j=1}^K P(e_j \mid t_{j-1}, n_{j-1}) \right] \exp(-\eta_{K+1}(t_{K+1} - t_K)),$$

where $\eta_j = (n_{j-1} + 1)\lambda + f(n_{j-1})\mu$ is the total intensity of indel rates across all positions and $f(x) = \sum_{k=1}^x (x - k + 1)d(k)$ is used to sum deletion probabilities over all positions and all allowable deletion sizes. The probability density of each event involves choosing the time given the current time and length, the type of event (insertion or deletion) given the current length and that the event occurs, and the position and size of the event given its type and the current length.

$$P(e_j \mid t_{j-1}, n_{j-1}) = P(p_j, l_j \mid id_j, n_{j-1})P(id_j \mid n_{j-1})P(t_j \mid t_{j-1}, n_{j-1}),$$

where $P(t_j \mid t_{j-1}, n_{j-1}) = \eta_j \exp(-\eta_j(t_j - t_{j-1}))$ and

$$P(id_j \mid n_{j-1}) = \begin{cases} \frac{(n_{j-1}+1)\lambda}{\eta_j} & \text{if } id_j = \text{in} \\ \frac{f(n_{j-1})\mu}{\eta_j} & \text{if } id_j = \text{del}. \end{cases}$$

If $id_j = \text{in}$, for $p_j \in \{0, 1, \dots, n_{j-1}\}$ and $l_j \in \{1, 2, \dots\}$, then $P(p_j, l_j \mid id_j, n_{j-1}) = \frac{i(l_j)}{n_{j-1}+1}$. If $id_j = \text{del}$, for $p_j \in \{0, 1, \dots, n_{j-1} - 1\}$ and $l_j \in \{1, \dots, n_{j-1} - p_j\}$, then $P(p_j, l_j \mid id_j, n_{j-1}) = \frac{d(l_j)}{f(n_{j-1})}$. Putting this together,

$$P(e_j \mid t_{j-1}, n_{j-1}) = \begin{cases} \exp(-\eta_j(t_j - t_{j-1}))\lambda i(l_j) & \text{if } id_j = \text{in} \\ \exp(-\eta_j(t_j - t_{j-1}))\mu d(l_j) & \text{if } id_j = \text{del}. \end{cases}$$

Therefore, the probability density for an indel history on a single edge simplifies to

$$P(h \mid v, n_0) = \exp \left(- \sum_{j=1}^{K+1} \eta_j(t_j - t_{j-1}) \right) \prod_{j=1}^K \left[(\lambda i(l_j))^{I_{\{id_j=\text{in}\}}} (\mu d(l_j))^{I_{\{id_j=\text{del}\}}} \right]. \quad (1)$$

On the tree For a given tree $T = (\tau, V)$ with n taxa, let $V = (v_1, \dots, v_{2n-3})$ and $H = (h_1, \dots, h_{2n-3})$ refer to the branch lengths and the indel histories of the edges. For convenience in calculation of the probability of an indel history on a tree, we assume one node of the tree T is a root, and then define an artificial parent node for each edge be the node nearest to this root. Let h_j represent the indel history on the j th edge under the defined direction and n_j denote the sequence length of its parent node. Then, the probability density for the indel history H on the tree T is $P(H \mid T) = q(n_r) \prod_{j=1}^{2n-3} P(h_j \mid v_j, n_j)$, where n_r indicates the sequence length at the root and $q(\cdot)$ represents the equilibrium length distribution. The probability density $P(h_j \mid v_j, n_j)$ is calculated by formula (1) for the j th edge.

2.4 Specification of the prior distributions

The parameters are partitioned into three parts, each with an independent prior distribution.

Branch lengths $\Theta_{\text{tree}} = \gamma$: For a prior on γ , we assume the density $g(\gamma) = \frac{\alpha_\gamma}{(1+\alpha_\gamma\gamma)^2}$ for $\gamma > 0$, which arises as a ratio of independent exponential random variables and has a very heavy tail; the mean is infinite, but the median is $1/\alpha_\gamma$.

Substitution model $\Theta_{\text{sub}} = (\pi, \kappa)$: We assume a Dirichlet prior distribution with parameters $\alpha_\pi = (\alpha_A, \alpha_C, \alpha_G, \alpha_T)$ for π and a ratio of exponentials prior distribution with a parameter α_κ for κ .

Indel model $\Theta_{\text{ID}} = (r, r_d, \lambda)$: We assume a beta prior distribution with parameters (α_r, β_r) and $(\alpha_{r_d}, \beta_{r_d})$ for r and r_d , respectively, and an exponential prior distribution with a parameter α_λ for λ .

3 MCMC approach

We sample from the joint posterior distribution $P(H, \tau, V, \Theta \mid S)$ using MCMC to estimate the alignment, tree, and model parameters and to quantify uncertainty in these estimates. To sample from the entire state space containing a tree T , an indel history H on the tree, and model parameters Θ , we use several MCMC updates employing a random-scan line (Liu *et al.*, 1995), Metropolis-within-Gibbs (Tierney, 1994) approach. MCMC

using reversible jump (Green, 1995) is adopted in updates involving the indel history due to changes in the dimension of the state space. Our MCMC proposal methods have four categories (overview of the proposal methods in these four categories is shown in Web Appendix E). The proposal in the first category updates the branch length (V) of a randomly selected edge. Although the times of the indel events (H) on the edge change in proportion to the change of the edge length, this update method does not vary alignments. Proposal methods in the second category select an edge of the tree at random and propose a new indel history (H) on the edge, conditional on the fixed sequence lengths of the two nodes connected by the edge. Here, the proposed new history can modify the alignment of sequences (A). Proposal methods in the third category pick an internal node and update an indel history (H) on three edges, which are adjacent to the internal node. This method updates an alignment (A), a sequence length at the internal node, and branch lengths (V) of the edges adjacent to the internal node. The last category contains proposal methods of subtree pruning and regrafting which update a tree topology (τ), a sequence length at an internal node, an indel history (H), an alignment (A), and the collection of branch lengths (V).

As part of the effort made to validate the implementation of our MCMC methods, we generate many data sets from the prior distribution, run MCMC on each one, calculate summary statistics of interest from each sample, and average these across samples. Close agreement between these results and expected values from the prior distribution is evidence of correct derivation and implementation of our MCMC approach (see Web Appendix F for detailed procedure and results).

Shim (2010) describes all the update methods in detail. The proposals for changing the tree and substitution model parameters are common in the Bayesian phylogenetics literature. The proposals that modify an indel history on an edge are novel to the modeling approach in this paper and are incorporated into the proposals that modify other parts of the parameter space as well. Here, we describe in detail algorithms for proposing an indel history on a single edge, conditional on the sequence lengths of the two nodes connected by the edge being fixed.

3.1 Propose a new indel history on a single edge

For a given edge of length v with parent and child nodes of sequence lengths n_0 and n_v , respectively, we propose new indel history h . A provisional history is generated sequentially starting from the sequence at the parent node using the Markov model for the indel process. The time to the next event is generated from an exponential distribution whose rate is the total sum of the rates of all possible next events on the current sequence. An insertion or deletion event is proposed according to its rate on the current sequence. The sequence length changes after each indel event. This process proceeds until the next event time exceeds the length of the edge. If the length of the final sequence differs from n_v , one additional event is appended to the provisional history at a random time between

the last event and v with the type and fragment size chosen to match the required sequence length at the end. The detailed proposal algorithm is provided in Web Appendix G.

The probability of proposing a indel history $h = (e_1, e_2, \dots, e_K)$ under this procedure, $Q(h \mid v, n_0, n_v)$, is calculated as follows. If $K > 0$,

$$Q(h \mid v, n_0, n_v) = \left[\prod_{i=1}^{K-1} P(e_i \mid t_{i-1}, n_{i-1}) \right] P(e_K \mid t_{K-1}, n_{K-1}, v, n_v).$$

Define η_i and $f(x)$ as above and let $q_{\text{in}}(\cdot)$ and $q_{\text{del}}(\cdot)$ be the probabilities of proposing an insertion and a deletion of a given size, respectively. Then,

$$P(e_i \mid t_{i-1}, n_{i-1}) = \begin{cases} \exp(-\eta_i(t_i - t_{i-1})) \lambda q_{\text{in}}(l_i) & \text{if } id_i = \text{in} \\ \exp(-\eta_i(t_i - t_{i-1})) \frac{f(n_{i-1}) \mu q_{\text{del}}(l_i)}{n_{i-1} - l_i + 1} & \text{if } id_i = \text{del} \end{cases}$$

and $P(e_K \mid t_{K-1}, n_{K-1}, v, n_v)$ is

$$\begin{cases} \exp(-\eta_K(t_K - t_{K-1}) - \eta_{K+1}(v - t_K)) \lambda q_{\text{in}}(l_K) + \frac{\exp(-\eta_K(v - t_{K-1}))}{(v - t_{K-1})(n_{K-1} + 1)} & \text{if } id_i = \text{in} \\ \exp(-\eta_K(t_K - t_{K-1}) - \eta_{K+1}(v - t_K)) \frac{f(n_{K-1}) \mu q_{\text{del}}(l_K)}{n_{K-1} - l_K + 1} + \frac{\exp(-\eta_K(v - t_{K-1}))}{(v - t_{K-1})(n_{K-1} - l_K + 1)} & \text{if } id_i = \text{del}. \end{cases}$$

If there are no events ($K = 0$), then $Q(h \mid v, n_0, n_v) = \exp(-\eta_1 v)$.

3.1.1 Propose a new indel history on a single edge considering the sequence length at the child node

The proposal introduced above takes into account the sequence length at the child node (n_v) only at the last step when proposing one additional event. This can lead to a high probability of proposing unlikely histories that are longer than more likely histories. For instance, if a sequence at a parent node has two more bases than a sequence at a child node, a single deletion event with a fragment size of two bases might be the most probable history. However, about half of the proposed histories will begin with an insertion event, and a further deletion event will be required. An alternative proposal method includes these modifications: (1) an increased probability of proposing no additional events when the current sequence length matches the target; (2) an increased probability of proposing an insertion (deletion) when the target length is greater (less) than the current length; and (3) an increased probability of proposing a fragment size to match the target sequence length. Although this proposal introduces a number of tuning parameters and comparison steps, we observe that it helps to increase MCMC mixing. The detailed description and proposal probability are provided in Web Appendix H.

3.2 Alignment summary

To summarize samples of alignments, we present an alignment with maximal expected accuracy and visualize uncertainty for every column and character of the alignment with

color, which is accomplished using the method proposed by Bradley *et al.* (2009) and implemented in the program FSA (Fast Statistical Alignment).

The software FSA consists of two separate parts. The first part of FSA performs pair-wise comparisons of the input sequences to estimate the posterior probabilities that individual characters are aligned using the standard three- or five-state pair hidden Markov model (Durbin *et al.*, 1998). The second part of FSA constructs a multiple alignment from the posterior probabilities estimated at the first part using a sequence annealing technique (Schwartz and Pachter, 2007). This procedure produces the multiple alignment with maximal expected accuracy, which is defined as a multiple alignment with minimal expected distance to the true alignment. The true alignment is treated as a random variable whose distribution is determined under a statistical model used in the first step.

Instead of the first step of FSA, we estimate the posterior probabilities for each pair of sequences from our multiple alignment samples. Then, we adopt the second part of FSA to construct the multiple alignment with maximal expected accuracy. Since the posterior probabilities used in the second step of FSA are estimated under our model, the final multiple alignment has maximal expected accuracy under our model (see Web Appendix I). Figure 2 shows an example of alignment summarization. Each character (gap) is colored according to the expected accuracy with which each character (gap) is aligned to other characters or gaps in the column. We note that FSA allows alignment uncertainty to be evaluated by other measurements : sensitivity, specificity, certainty, and consistency.

4 Applications

We apply our approach to joint estimation of the alignment and tree (BayesCAT) to a data set from Redelings and Suchard (2005), and then compare the performance of BayesCAT with the traditional sequential methods and with an alternative joint model approach, BAli-Phy (Suchard and Redelings, 2006). In addition, we conducted a comparison on simulated data, where the true tree, indel history, and alignment are known, and provide a detailed procedure and comparison results in Web Appendix J.

4.1 Data description: 5S rRNA

A question of interest in Redelings and Suchard (2005) is whether the Archaea form a monophyletic group, one of the important unresolved question about deep branches in the Tree of Life (Brown and Doolittle, 1997).

We begin with a brief summary of the background introduced in Redelings and Suchard (2005). The division of all living organisms into the three domains (Archaea, Bacteria, and Eucarya) was proposed by Woese *et al.* (1990) and has been supported by research into the molecular biology of Archaea (Brown and Doolittle, 1997). Woese *et al.* (1990)

Table 1: **5S rRNA : Data description.** Note that boldface here and in other tables represents Archaea species. Abbreviation corresponding to each species is shown in parentheses.

Taxa	Domain	Order
Escherichia coli (EC)	Bacteria	Proteobacteria
Homo sapiens (HS)	Eukaryotes	Metazoa
Halobacterium salinarum (HA)	Archaea	Euryarchaeota
Pyrococcus woesei (PW)	Archaea	Euryarchaeota
Sulfolobus acidocaldarius (SA)	Archaea	Crenarchaeota

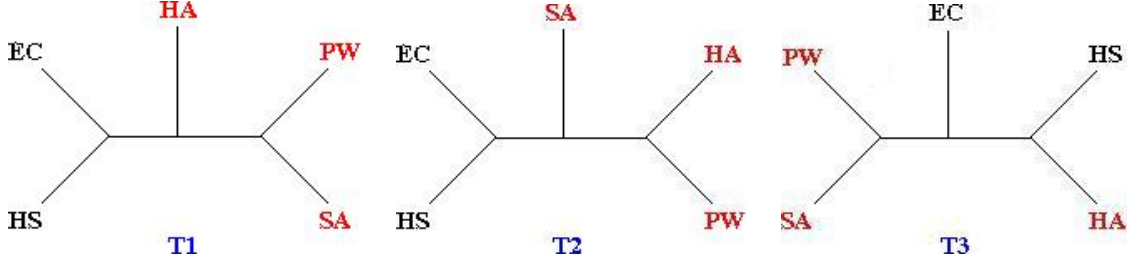
suggest that Archaea form a monophyletic group, but some other analyses suggest that the Crenarchaeotes, also called Eocytes after Lake (1991), separated from the remaining Archaea and form a clade with the Eukaryotes (Rivera and Lake, 1992). These conflicting results suggest two alternative hypotheses about the early branching in the Tree of Life. The archaea tree represents the hypothesis that the Archaea form a monophyletic group while the eocyte tree denotes the alternative hypothesis that the eocyte Archaea are more closely related to Eukaryotes than to the remaining Archaea (See Figure 5 in Redelings and Suchard (2005)).

The 5S rRNA, a component of the large ribosomal subunit, is found in Archaea, Bacteria, and Eukaryotes and has a highly conserved secondary structure (Barciszewska *et al.*, 2001). The 5S rRNA sequences in the data set range in length from 120 to 126 base pairs and the lowest pairwise sequence identity is around 46% (Redelings and Suchard, 2005). Table 1 lists five taxa used in the analysis. Redelings and Suchard (2005) also used this data to compare BALi-Phy to the traditional sequential approach.

4.2 Model and prior distributions

We use the HKY model (Hasegawa *et al.*, 1985) in our analysis here (see Web Appendix B for details of the HKY model), and we use the geometric distribution for the deletion fragment size. The prior distribution of the remaining parameters are described in Section 2.4. We assume a Dirichlet prior distribution with parameters $\alpha_\pi = (13.3, 21.7, 23.1, 11.9)$ for π . The parameter α_π is selected to have the observed frequencies of bases as a mean and to cover broad regions. We assume a beta prior distribution with parameters (100, 12200) and (3, 15) for r and r_d , respectively. These prior distributions are selected to cover reasonably broad regions based on the observed sequence lengths. The posterior estimates of each parameter, together with a prior mean (and median for γ and κ), are summarized in Web Table 1.

Table 2: **5S rRNA : Summary of posterior distributions of the topology.** $T1$, $T2$, and $T3$ are the top three topologies ranked by posterior probabilities from BayesCAT. The full name corresponding to each abbreviation can be found in Table 1. Archaea taxa are shown in boldface.



Method	$T1$	$T2$	$T3$	EC,HS HA,PW,SA
BayesCAT	0.205	0.170	0.130	0.414
BAlI-Phy	0.284	0.103	0.189	0.418
MrBayes+ClustalW	0.700	0.172	<0.001	>0.999

4.3 Phylogeny estimation

Table 2 shows posterior probabilities of the top three topologies, ranked by posterior probabilities from BayesCAT. No single topology is strongly supported by BayesCAT as the support for the most probable topology is only 0.205. The posterior probabilities for the archaea tree ($T2$) and the eocyte tree (not shown) are 0.17 and 0.078, respectively. The split, which supports the hypothesis of archaeal monophyly, has a posterior probability of 0.414 (Table 2).

Table 2 also lists posterior probabilities from the traditional sequential approach where we apply MrBayes to the alignment determined using ClustalW. The most probable topology ($T1$) has a high support (0.7). The archaea tree ($T2$) has a posterior probability of 0.172 while the support for the eocyte tree is less than 0.001 (not shown). Unlike BayesCAT, the traditional sequential approach strongly supports the split for archaeal monophyly (posterior probability > 0.999).

To compare BayesCAT to another joint model, posterior probabilities from BAlI-Phy are also presented in Table 2. We note that these posterior probabilities are not identical to results in Redelings and Suchard (2005) as the version of BAlI-Phy used here improves upon the version used in their publication. Although BAlI-Phy assumes a different indel model, its results are very similar to that from BayesCAT. The top three topologies ranked by posterior probabilities from BAlI-Phy also are $T1$, $T2$ and $T3$, although $T3$ has higher posterior probability than $T2$ with the BAlI-Phy model. Like BayesCAT, no single

topology is strongly supported by BALi-Phy. Of particular note, the posterior probabilities of the hypothesis of archaeal monophyly in BALi-Phy (0.418) and BayesCAT (0.414) are very similar.

In summary, the traditional sequential analysis based on the 5S rRNA sequences leads to strong posterior support for a single topology consistent with the hypothesis of archaeal monophyly, but the joint model moderates this strong support by considering alignment uncertainty. In addition, we observe that two different joint models yield quite similar support for the hypothesis of archaeal monophyly.

4.4 Summary of alignment samples

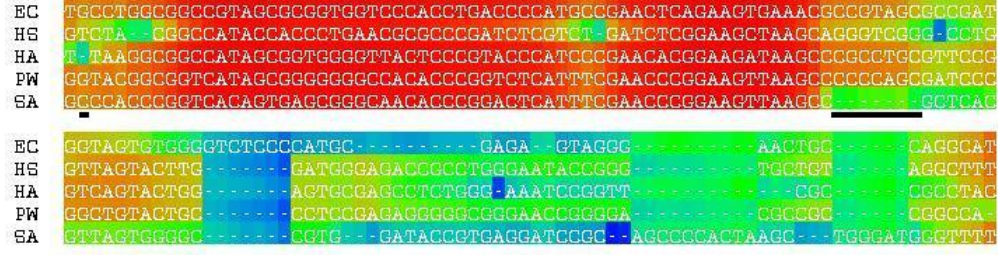
Alignment samples from BayesCAT and BALi-Phy are summarized in Figure 2 using the procedure described in Section 3.2. Although BALi-Phy provides its own summarization method, we use the same summarization procedure for both programs to focus the comparison on the alignment distributions and not the summarization methods. The two point estimates under the different joint models have the same columns in the first half of the alignment except for positioning of two gaps (underlined in Figure 2). In addition, red color in the first half of the columns indicates that the two point estimates have high expected accuracy under each model. In contrast, in the second half of the alignment, the two point estimates are quite different and also show low expected accuracy, illustrated by the blue color. Most of the gaps observed in both alignments are not shared by multiple taxa, which is consistent with the explanation that most of the indel events happen on the external edges (Section 4.5).

To investigate differences in the alignment distribution between BayesCAT and BALi-Phy, we plot the pairwise homology posterior probabilities from each method (see Web Figure 3 (a)). Points form a broad band around a diagonal, but no point is substantially far from the diagonal. To compare with variability from Monte Carlo error, we plot the pairwise homology probabilities from two different MCMC samples using BayesCAT (see Web Figure 3 (b)). As the two plots show a similar deviation from the diagonal, the difference in the alignment distribution samples between the two methods may be due in large part to Monte Carlo error.

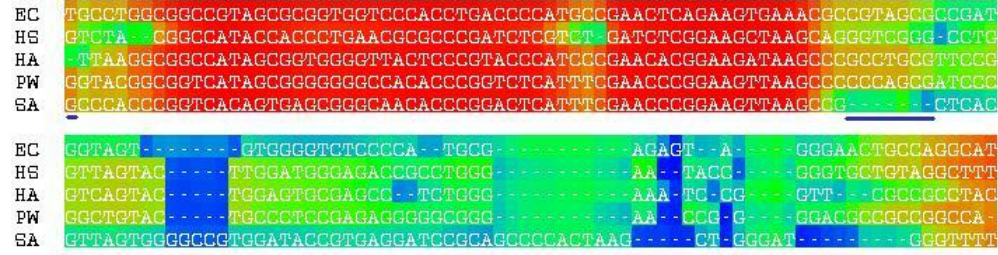
The multiple sequence alignment estimated using FSA is given in Web Figure 4. The beginning of the alignment, colored in red, is very similar to that of point estimates under joint models in terms of an alignment of residues as well as their expected accuracy, but very little similarity is observed in the remaining part.

4.5 Information on the indel process

Since we model indel events directly, some information about the indel process can be inferred by using our approach and not by BALi-Phy. Web Figure 5 shows the posterior estimate of realized indel fragment size distribution, which is obtained by first collecting



(a) BayesCAT



(b) BALi-Phy



Figure 2: **5S rRNA : Summary of alignment samples.** Alignment samples from BayesCAT (a) and BALi-Phy (b) are summarized using the procedure described in Section 3.2

empirical indel fragment size distributions from each sample, and then averaging over all samples. This distribution has modes at sizes one and seven.

Another quantity we can estimate is the number of indel events on each split. Table 3 shows the posterior mean of number of indel events given occurrence of each split. Most splits corresponding to external edges (the first five splits in Table 3) include more than one indel event while the mean number of indel events on internal edges are fewer than one. Occurrence of indel events on internal edges implies that two or three leaves (in the five-taxon case) share these events. Thus, the fact that most of the indel events are observed on the external edges supports that the 5S rRNA sequences we examined do not have a strong phylogenetic signal for shared indel events.

To investigate whether the expected number of indel events vary with branch length, we also list the posterior mean of edge length given occurrence of each split in the fourth column of Table 3. Edges with more than one indel event are longer than the remaining edges. The second split has the largest number of indel events (3.5) on the longest edge

Table 3: **5S rRNA : Posterior mean number of indel events on each split (BayesCAT)**. The second column lists posterior probabilities for each split. The posterior means of the number of indel events and the edge length given occurrence of each split are shown in the third and fourth columns, respectively. Archaea taxa are shown in boldface.

Split	PP of split	# of indels	edge length
EC HS, HA , PW , SA	1	2.3	0.456
HS EC, HA , PW , SA	1	3.5	0.464
HA EC,HS, PW , SA	1	2.7	0.264
PW EC,HS, HA , SA	1	0.8	0.147
SA EC,HS, HA , PW	1	3.3	0.366
EC,HS HA , PW , SA	0.41	0.16	0.112
EC, HA HS, PW , SA	0.08	0.33	0.046
EC, PW HS, HA , SA	0.12	0.11	0.080
EC, SA HS, HA , PW	0.16	0.55	0.076
HS, HA EC, PW , SA	0.35	0.38	0.112
HS, PW EC, HA , SA	0.007	0	0.025
HS, SA EC, HA , PW	0.161	0.37	0.097
HA , PW EC,HS, SA	0.303	0.04	0.090
HA , SA EC,HS, PW	0.053	0.26	0.044
PW , SA EC,HS, HA	0.38	0.41	0.105

length (0.464) while the split with no indel events has the shortest edge length (0.025).

4.6 Convergence

We run three MCMC chains from different starting points. Each run has 1,000,000 iterations and we sampled every 1000 iterations. To assess convergence for continuous parameters, we compute Gelman-Rubin R statistics (Gelman and Rubin, 1992) for sampled external branch lengths and substitution and indel parameters. All statistics are less than 1.05 which is consistent with convergence. Convergence for the tree topology is evaluated as follows. For each clade which appeared in any of the three runs, we calculate the relative frequency with which the clade occurs in each of the runs. Differences between the minimum and maximum values of these relative frequencies over three runs are less than 5% for all clades.

5 Discussion

We have developed a joint model for co-estimation of the alignment and tree. Our general indel model allows arbitrary-length overlapping indels and a general distribution for the indel fragment size. We designed and implemented MCMC methods to carry out Bayesian inference of multiple sequence alignment and phylogeny on the basis of this model. Our method for joint estimation improves estimates from the traditional sequential approach by accounting for uncertainty in the alignment in phylogeny inferences, which is demonstrated by real data and simulated data (results from simulated data is given in Web Appendix J).

Our method is the first approach which includes a complete history of indel events mapped onto the tree as the state space in the Markov Chain. A large state space containing the complete history of indel events makes our MCMC approach more challenging, but it enables us to infer more information about the indel process itself than can be done with alternative joint model approaches. Inferred information about the indel process has the potential to be very valuable for some questions of biological interest for some data sets. In addition to quantities presented in this paper, we can infer more information, e.g., positions of indel events and the proportion of overlapping indels. Our method would be useful to a biologist interested in the indel process itself.

The alternative approaches sum over all possible indel histories, which places severe constraints on the choice of indel models, e.g., distribution for indel fragment size and number of indel events. Thus, our method has the advantage of being relatively easy to extend to model more closely real processes of insertion and deletion.

To summarize alignment samples, we suggest using a method implemented in FSA and describe how to use it in the joint estimation setting. Although this suggestion improves on alternative summarization methods in terms of estimating a point-estimate and showing uncertainty in the point-estimate, the information provided by this summary is still limited. We also investigate pairwise posterior probabilities of homology, but our investigation is still ad hoc and cannot provide enough information to fully summarize alignment uncertainty. Thus, developing methods to present alignment distribution in a more informative manner remains an interesting open challenge.

Although our method has several advantages relative to the alternative approaches, it still has points which need to be investigated more thoroughly. We have not observed notable differences in inferences between our method and BALi-Phy in data analysis although our method assumes a more general indel model. We need to investigate when our method can have some advantages. Such advantages would presumably be most likely to occur if the true history is likely to contain overlapping indel events.

Acknowledgements

We thank Cécile Ané, Colin Dewey, David Baum, and Michael Newton for helpful suggestions. We thank the authors of FSA for modifying their software for our use, and the authors of BAli-Phy for sharing their research experience in joint estimation of alignment and tree.

Supplementary Materials

Web Appendix, Web Figure, and Web Table are available with this paper at <https://github.com/heejungshim/BayesCAT/tree/master/doc/paper>.

References

- Barciszewska, M. Z., Szymański, M., Erdmann, V. A. and Barciszewski, J. (2001) Structure and functions of 5S rRNA. *Acta Biochimica Polonica*, **48**, 191–8.
- Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I. and Pachter, L. (2009) Fast Statistical Alignment. *PLoS Computational Biology*, **5**, e1000392.
- Brown, J. R. and Doolittle, W. F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiology and molecular biology reviews*, **61**, 456–502.
- Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) Biological sequence analysis: Probabilistic models of proteins and nucleic acids. *Cambridge University Press, Cambridge, UK*.
- Gelman, A. and Rubin, D. B. (1992) Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, **7**, 457–472.
- Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
- Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.
- Lake, J. A. (1991) The order of sequence alignment can bias the selection of tree topology. *Molecular Biology and Evolution*, **8**, 378–85.
- Liu, J. S., Wong, W. H. and Kong, A. (1995) Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 157–169.

- Liu, K., Raghavan, S., Nelesen, S., Linder, C. R. and Warnow, T. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, **324**, 1561–1564.
- Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P. and Linder, C. R. (2012) SATE-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic biology*, **61**, 90–106. URL <http://sysbio.oxfordjournals.org/content/61/1/90>.
- Lunter, G., Miklós, I., Drummond, A., Jensen, J. L. and Hein, J. (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**.
- Lutzoni, F., Wagner, P., Reeb, V. and Zoller, S. (2000) Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Systematic Biology*, **49**, 628–51.
- Miklós, I., Lunter, G. A. and Holmes, I. (2004) A "Long Indel" Model For Evolutionary Sequence Alignment. *Molecular Biology and Evolution*, **21**, 529–540.
- Nelesen, S., Liu, K., Zhao, D., Linder, C. R. and Warnow, T. (2008) The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25–36.
- Novák, A., Miklós, I., Lyngsø, R. and Hein, J. (2008) StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics (Oxford, England)*, **24**, 2403–4. URL <http://bioinformatics.oxfordjournals.org/content/24/20/2403>.
- Redelings, B. D. and Suchard, M. A. (2005) Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology*, **54**, 401–418.
- Redelings, B. D. and Suchard, M. A. (2007) Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evolutionary Biology*, **7**.
- Rivera, M. C. and Lake, J. A. (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science*, **257**, 74–6.
- Schwartz, A. S. and Pachter, L. (2007) Multiple alignment by sequence annealing. *Bioinformatics*, **23**, e24–9.
- Shim, H. (2010) Bayescat : Bayesian co-estimation of alignment and tree. *PhD Thesis, Department of Statistics, University of Wisconsin at Madison*.
- Sinsheimer, J. S. (1994) Extensions to evolutionary parsimony. *Ph.D. thesis, University of California, Los Angeles*.

- Suchard, M. A. and Redelings, B. D. (2006) BALi-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, **22**, 2047–2048.
- Thorne, J. L. and Kishino, H. (1992) Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution*, **9**, 1148–62.
- Thorne, J. L., Kishino, H. and Felsenstein, J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, **33**, 114–124.
- Thorne, J. L., Kishino, H. and Felsenstein, J. (1992) Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, **34**, 3–16.
- Tierney, L. (1994) Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, **22**, 1701–1728.
- Varón, A., Vinh, L. S. and Wheeler, W. C. (2010) POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics*, **26**, 72–85. URL <http://doi.wiley.com/10.1111/j.1096-0031.2009.00282.x>.
- Woese, C. R., Kandler, O. and Wheelis, M. L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 4576–9.
- Wong, K. M., Suchard, M. A. and Huelsenbeck, J. P. (2008) Alignment Uncertainty and Genomic Analysis. *Science*, **319**, 473–476.